# Quiz 01

Sep 26, 2017.

1. (30 pts) How many bits does it take to store a binary floating point number of the form $\pm 1.a_1a_2\cdots a_t \times 2^e$ with $t = 10$, $a_j \in \{0,1\}$, $-14 \le e \le 15$? Write down the binary floating number representation (binary machine number, a finite sequence of 0, 1) of $-0.6875$. Explain. Leave some spacing between sign and exponent and between exponent and mantissa for easy reading.

   **Ans**:

   There are total 30 different exponents ($-14 \le e \le 15$). It takes 5 bits to give 30 or more different exponents ($2^5 = 32$). Total bits $= 1 + 10 + 5 = 16$ **(15pts)**.

   The range of the 5-bit binary exponent $c = (b_1b_2b_3b_4b_5)_2$, $b_i = 0, 1$, is $0 \le c \le 31$. In order to cover the range $-14 \le e \le 15$, one should take $e = c - 15$, so that $e = -15$ and $e = 16$ can be reserved for underflow and overflow, respectively. With $e = c - 15$, the binary machine number is given by:

   $$-0.6875 = -(1.011)_2 \times 2^{-1} = 1\ \ 01110\ \ 0110000000\ \ \textbf{(15 pts)}$$

2. (30 pts) <u>Derive</u> an upper bound for relative error caused by chopping for the floating point system in problem 1 (also known as $\varepsilon_M$). Give an upper bound in terms of $\varepsilon_M$ on the relative error of evaluating $x \times y$ with the floating point arithmetics.

   **Ans**:

   $$\begin{aligned} \frac{|x - fl_{chop}(x)|}{|x|} &= \left|\frac{0.0\ldots0a_{t+1}\ldots \times 2^e}{1.a_1\ldots a_{t+1}\ldots \times 2^e}\right| \\ &= \left|\frac{0.a_{t+1}\ldots}{1.a_1\ldots a_{t+1}\ldots}\right| \times 2^{-t} \\ &\le \left|\frac{1}{1}\right| \times 2^{-t} = 2^{-t} = 2^{-10}\ \ \textbf{(15 pts)} \end{aligned}$$

   $$\begin{aligned} \frac{|x \times y - fl(fl(x) \otimes fl(y))|}{|x \times y|} &= \left|\frac{x \times y - (x(1+\delta_1) \times y(1+\delta_2))(1+\delta_3)}{x \times y}\right| \\ &\approx \left|\frac{x \times y - x \times y(1+\delta_1+\delta_2+\delta_3)}{x \times y}\right| \\ &\le |\delta_1| + |\delta_2| + |\delta_3| \le 3\epsilon_M\ \ \textbf{(15 pts)} \end{aligned}$$

3. (20 pts) Solve for $x^2 - 2100x + 1 = 0$ to 15 correct digits. Explain how you find your answer (direct evaluation using 'calculator' will receive no credits).

   **Ans**:

   $x_1 = \frac{2100+\sqrt{2100^2-4}}{2} = 2.09999952380942e+03$ **(5 pts)**
   $x_2 = 1/x_1$ or $\frac{2}{2100+\sqrt{2100^2-4}} = 4.76190584170225e-04$ **(5 pts)**
   Code **(10 pts)**
   Extra points by writting C **(2 pts)**

4. (20 pts) Find the smallest $N$ so that $\left|\sum_{i=0}^{N} \frac{3^i}{i!} - e^3\right| < 10^{-5}$. Let your code print the answer $N$ and $\left|\sum_{i=0}^{N} \frac{3^i}{i!} - e^3\right|$ on screen, and also write them down on the answer sheet. Extra credits for more efficient method(s).

   **Ans**:

   N=15 **(5 pts)**
   absolute error = 2.49221685777457e-06 ("format short" is ok.) **(5 pts)**
   Code **(10 pts)**
   Extra points by writting C **(2 pts)**
   Extra points by nested summation **(4 pts)**

Name your codes in the same format as 104000001_p03.m or 103000002_p04.c .

Numerical Analysis I, Fall 2017 (http://www.math.nthu.edu.tw/~wangwc/)

# Quiz 01

Sep 26, 2017.

1. (30 pts) How many bits does it take to store a binary floating point number of the form $\pm 1.a_1 a_2 \cdots a_t \times 2^e$ with $t = 10$, $a_j \in \{0, 1\}$, $-14 \leq e \leq 15$? Write down the binary floating number representation (binary machine number, a finite sequence of 0, 1) of $-0.6875$. Explain. Leave some spacing between sign and exponent and between exponent and mantissa for easy reading.

2. (30 pts) Derive an upper bound for relative error caused by chopping for the floating point system in problem 1 (also known as $\varepsilon_M$). Give an upper bound in terms of $\varepsilon_M$ on the relative error of evaluating $x \times y$ with the floating point arithmetics.

3. (20 pts) Solve for $x^2 - 2100x + 1 = 0$ to 15 correct digits. Explain how you find your answer (direct evaluation using 'calculator' will receive no credits).

4. (20 pts) Find the smallest $N$ so that $\left|\sum_{i=0}^{N} \frac{3^i}{i!} - e^3\right| < 10^{-5}$. Let your code print the answer $N$ and $\left|\sum_{i=0}^{N} \frac{3^i}{i!} - e^3\right|$ on screen, and also write them down on the answer sheet. Extra credits for more efficient method(s).

Name your codes in the same format as 104000001_p03.m or 103000002_p04.c .